International Journal of Advanced Education and Research
www.alleducationjournal.com
ISSN: 2455-5746
Received: 21-12-2021, Accepted: 06-01-2022, Published: 22-01-2022
Volume 7, Issue 1, 2022, Page No. 1-5

Application of item response theory in calibrating 2020 neco mathematics multiple-choice questions

Romy O Okoye, Somtoo Victor Nduka
Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

Abstract

Item analysis is indispensable in the field of measurement and evaluation, as it provides a means of assessing the quality of test items. This study demonstrates the use of the item response theory to calibrate the test item parameters (item difficulty and item discrimination) of the 2020 NECO Mathematics multiple choice questions. The population for the study was made up of all the 26,421 SSIII students of the 2020/2021 academic session who enrolled for the NECO examinations in Anambra State, out of which a sample of 1,828 was selected by the use of purposive sampling technique. The students' responses were analyzed, and the 4PLM was used to calibrate the IRT-based parameters. The findings from this study revealed that out of the sixty (60) items in the test, twenty eight (28) were moderately easy and twenty nine (29) were moderately difficult as they respectively fell between -3 to 0 and 0 to +3 on the logit scale. These fifty seven (57) items were, therefore, considered as good items. However, three (3) items were found to be too easy as they fell below -3 on the logit scale, and therefore, were considered bad items. In terms of the discriminating powers, eighteen (18) items had moderate discriminating power, nineteen (19) and seventeen (17) items respectively had high and very high discriminating powers. These fifty four (54) items were, therefore, considered as good items. However, two (2) and four (4) items respectively had low and very low discriminating power, and therefore, considered as bad items. The study concluded that the two item statistics (item difficulty and item discrimination) produced almost the same item characteristics, and therefore, recommended that item analysis should be maintained in the test development process because of its importance in determining the overall effectiveness of the test.

Keywords: item analysis, item response theory, item difficulty, item discrimination

Introduction

Education in its truest sense is an indispensable instrument for the overall development of any nation. The national policy on education (Federal Republic of Nigeria, 2014) [6] identified education as fundamental to the nation's developmental goals, and vital for the promotion of a progressive and united Nigeria.

To realize this goal of general development of the Nigerian society, there is need to maintain high standard and quality education delivery across the nation. The resulting effect of this educational process is quantified through measurement and evaluation.

Evaluation adds the quality of value judgement to measurement, which according to Omorogiuwa (2010) [17] can be defined as the assigning of numbers on a student's performance to show the extent to which a trait is present or absent, according to specified rules. Evaluation as stated by Olabode and Adeleke (2015) [16] is concerned with the quality of the curriculum, facility, and performance of pupils, using various tools. One of such tools is test. Others, as noted by Okoye (2015) [14] include observations, rating scales, interviews among others.

In Nigeria, there are several examination bodies charged with the responsibility of providing quality assessment of the degree to which educational objectives have been achieved at different stages of the educational system. Unfortunately, over the years, the outcome of the assessments provided by these examination bodies indicate a persistent students' low level of achievement, especially in Mathematics and other science subjects.

This consistent poor performance of students in Mathematics continues to draw the attention of major stakeholders in education. Researchers have tried to identify the causes, and proffer solutions to this disheartening trend. For example, Agwagah cited in Onah (2015) [18] stated that there are many causes of poor performance of students in mathematics, which if well tackled would help the students improve on their performances in the subject. The causes were found to include; students' lack of interest and/or negative attitude towards mathematics; lack of qualified mathematics teachers; teachers' own negative attitude and/or incompetence in certain concepts; poor method of teaching applied by the teachers in the classroom; teachers' non-use of instructional materials in the teaching of mathematical concepts especially some that seem abstract. Other researchers proffered ways that can be adopted in order to improve students' level of achievement. Anigbo (2014) [2] suggested improved teaching methodology and students' preparation for examination; Adegoke cited in Adegoke (2013) [1] suggested the use of multimedia instruction; Mbugua, Kibet, Muthaa and Nkonke, (2012) [10] suggested the provision of proper staffing, teaching and learning materials, and improved curriculum. However, despite all these suggestions, very little improvement has been recorded. A perusal of the above causes and suggestions revealed to the present researchers of the emphasis placed on factors that make up student's characteristics. Interestingly, it is very imperative to note that performance in any given test does not solely depend on the student's characteristics, but also on the test item characteristics. Laying credence to this, Ashikhia (2010) [4] identified the nature of the test items as a prominent factor that affects student's performances. Test item characteristics are the inherent qualities of the test which determines the overall effectiveness of the test, and these qualities are determined through item analysis. Item analysis is a process of examining students' responses to each test item in order to measure the quality of the test items.

The inherent qualities of test items can be studied and evaluated from different perspectives and theories. These theories are referred to Test Theories, and they provide the framework to improve the overall quality of tests by identifying the parameters of item difficulty, item discrimination and the effectiveness of alternatives. In educational measurement, there are two main test theories or frameworks; the Classical Test Theory (CTT) and the Item Response Theory (IRT). These frameworks are based on different theoretical assumptions and use different statistical approaches. IRT is primarily interested in the item-level information. This is in contrast to the CTT which primarily focuses on test-level information. IRT models are often referred to as latent trait models. The term latent is used to emphasize that discrete item responses are taken to be observable manifestations of hypothetical traits, constructs or attributes not directly observed, but are manifest responses. In other words, IRT describes the relationship between an examinee's test performance and the latent traits assumed to underlie such performance (Hambleton, Robin, & Xing, 2000) [8]. IRT models are lauded (Ojerinde, 2013; Wang & Hanson, 2001; Wiberg, 2004) [13, 21, 22] for their ability to generate invariant estimates of item and person parameters. That is, theoretically, IRT ability estimates (θ) are item-free (would not change if different items were used) and the item statistics are person-free (would not change if different persons were used). There are four (4) IRT models based on the number of item parameters. These are one-parameter logistic (1PL) model, two-parameter logistic (2PL) model, three-parameter logistic (3PL) model and four-parameter logistic (4PL) model. The 1PL model has only the item difficulty parameter *(b)*, the 2PL model has item difficulty *(b)* and item discrimination *(a)* parameters, the 3PL model has item difficulty *(b)*, item discrimination *(a)*, and guessing *(c)* parameters and the 4PL model has item difficulty *(b)*, item discrimination *(a)*, item guessing *(c)* and upper asymptote *(d)* parameters. The choice of the IRT model to employ in a study is data dependent. It is, therefore, important to check for model-data fit before calibrating the item parameters using the framework of the IRT. -2Log likelihood value is commonly used to check the model data fit. The model with the smallest -2log likelihood value is the best fit (Thorpe & Favia, 2012). Given the indispensable nature of test item analysis in the field of measurement and evaluation, and the effect of test item parameters on student's academic achievement, there is a need to carry out this present study to analyse the 2020 NECO Mathematics test items using IRT framework to determine the psychometric properties of the test items with a view to identify problematic items.

Purpose of the Study

The purpose of the study is to generate the item parameter estimates of 2020 NECO Mathematics multiple choice questions using the Item Response Theory framework.

Specifically, the study did:

1. determine the extent to which the IRT assumptions are met by the 2020 NECO mathematics multiple-choice test items.
2. identify the IRT model that best fits the data of 2020 NECO mathematics multiple-choice test items.
3. determine the IRT-based item difficulty estimates of 2020 NECO mathematics multiple-choice test items.
4. determine the IRT-based item discrimination estimates of 2020 NECO mathematics multiple-choice test items.

Research Questions

The following research questions were raised to guide this study.

1. To what extent are the IRT assumptions met by the 2020 NECO mathematics multiple-choice test items?
2. What IRT model best fits the data of 2020 NECO mathematics multiple-choice test items?
3. What are the IRT-based item difficulty estimates of 2020 NECO mathematics multiple-choice test items?
4. What are the IRT-based item discrimination estimates of 2020 NECO mathematics multiple-choice test items?

Method

The descriptive survey design was adopted in this study. A sample of 1,828 out of a population of 26,421 Senior Secondary School three (SS3) students who enrolled for the 2020/2021 Senior School Certificate Examination in Anambra State was used in the study. The instrument for data collection was the May/June 2020 NECO Mathematics Paper III. This instrument was administered to the students of the sampled schools by the researcher, assisted by the Mathematics teachers of the schools, under strict examination conditions. The data obtained were analyzed for local independence and dimensionality test using the Yen's Q3 statistics and Dimtest statistics respectively. The best fit model for the calibration of the IRT item parameter was determined using the Akaike Information Criterion (AIC) and Likelihood Ratio Test (LogLik), and the item parameters were calibrated using Maximum likelihood estimation technique in Jmetrik software.

Presentation and Analysis Of Data

This section focuses on the analysis of the data obtained from the administration of the research instrument. The results of the analysis are presented in tables and discussed according to the research questions that formed the thrust of this study.

Research Question 1: To what extent are the IRT assumptions met by the 2020 NECO mathematics multiple-choice test items?

Table 1: Dimtest Statistics of 2020 NECO Mathematics Multiple-Choice Test Items

TL	TGbar	T	P-value
17.1871	3.9326	13.1887	0.0000

The result in Table 1 indicates that 2020 NECO mathematics multiple-choice test items were multidimensional since p-value is less than 0.05 level of significance. Furthermore, if the difference between the number of items in Partitioning Subtest (PT) and the Assessment Subtest (AT) in a test is significant, there is evidence of multidimensionality (Anyawale, Isaac-Oloniyo & Abayomi, 2020) [3]. For this particular study, the difference between the AT and PT as shown in Table 1 is significant (T=13.1887, p<0.05). This led to the conclusion that the AT items were dimensionally distinct from the remaining items in PT. Therefore, multidimensionality was manifest in 2020 NECO mathematics multiple-choice test items. Local independence assumption was investigated with the Yen's Q3 statistics. Based on this statistics, residuals for any pair of items should be uncorrelated, and generally close to zero. Residual correlations that are high indicate a violation of the local independence assumption, and this suggests that the pair of items have something more in common than the rest of the item set have in common with each other (Marais, 2013). For this study, using Yen's Q3 to screen items for local dependence, 75% item residual correlations were below absolute value of 0.2. This indicates that the local independence assumption of the IRT was not grossly violated.

Research Question 2: What IRT model best fits the data of 2020 NECO mathematics multiple-choice test items?

Table 2: IRT Model Best fit of 2020 NECO Mathematics Multiple-Choice Test Items

Model	AIC	-2Loglik
1PL	131533.5	131321.4
2PL	131534.5	131324.4
3PL	130648.8	130318.8
4PL	130630.4	130180.0

The examinee responses were subjected to full information item factor analysis, and compared using Akaike Information Criterion (AIC), and Likelihood Ratio Test (LogLik) in order to establish the best fit model that provided the information for the calibration of item parameters embedded in the test data. The result in Table 4 indicates that 4PLM has the smallest information criteria in terms of Akaike information criteria and -2Loglik. Therefore, the 4PLM was used to calibrate the IRT-based parameter estimates of 2020 NECO mathematics multiple-choice test items.

Research Question 3: What are the IRT-based item difficulty estimates of 2020 NECO mathematics multiple-choice test items?

Table 3: IRT-Based Item Difficulty Estimates of 2020 NECO Mathematics Multiple-Choice Test Items

Item	b	Item	b	Item	b	Item	b
1	-5.96	16	0.29	31	-0.23	46	0.09
2	-5.94	17	0.66	32	-0.28	47	0.17
3	-2.53	18	-0.72	33	-0.14	48	0.12
4	1.21	19	-0.18	34	0.24	49	0.61
5	-0.1	20	0.03	35	-0.11	50	0.22
6	-0.05	21	0.32	36	-0.14	51	0.38
7	-5.85	22	0.13	37	-0.15	52	0.56
8	0.08	23	-2.13	38	-0.25	53	0.03
9	-0.15	24	0.71	39	-0.01	54	-0.04
10	-0.15	25	-0.22	40	0.29	55	0.45
11	-0.19	26	-0.08	41	-0.51	56	0.11
12	-0.35	27	0.71	42	0.78	57	-0.12
13	-0.57	28	0.03	43	0.06	58	-0.18
14	-2.54	29	0.04	44	0.82	59	0.19
15	0.06	30	-0.04	45	0.47	60	-1.18

Table 3 indicates that twenty-eight (28) items, that is, Items 3, 5, 6, 7, 9, 10, 11, 12, 13, 14, 18, 19, 23, 25, 26, 30, 31, 32, 33, 35, 36, 37, 38, 39, 41, 54, 57, 58 and 60 fall between-3 and 0 on the logit scale, indicating moderately easy items, while twenty-nine (29) items, that is, Items 4, 8, 15, 17, 20, 21, 22, 24, 27, 28, 29, 34, 40,42, 43, 44, 45,46, 47, 48, 49, 50, 51, 52, 53, 55, 56 and 59 fall between 0 and +3 on the logit scale, indicating moderately difficult items. Three (3) items, that is, Items 1, 2 and 7 fall below -3 on the logit scale indicating very easy items. Therefore, under IRT-based item estimates of 2020 NECO mathematics multiple-choice test items, three (3) items were found to be too easy.

Research Question 4: What are the IRT-based item discrimination estimates of 2020 NECO mathematics multiple-choice test items?

Table 4: IRT-Based Item Discrimination Estimates of 2020 NECO Mathematics Multiple-Choice Test Items

Item	a	Item	a	Item	a	Item	a
1	1.35	16	0.91	31	1.40	46	2.09
2	0.32	17	2.51	32	1.84	47	2.03
3	0.51	18	2.32	33	1.71	48	1.55
4	0.54	19	1.56	34	1.70	49	2.61
5	1.51	20	1.29	35	1.67	50	1.76
6	1.24	21	1.29	36	1.85	51	1.60
7	0.21	22	1.42	37	1.28	52	2.69
8	1.39	23	0.37	38	1.53	53	1.40
9	1.20	24	2.44	39	1.75	54	1.71
10	1.62	25	1.26	40	1.26	55	2.51
11	1.30	26	1.18	41	1.20	56	1.61
12	1.40	27	2.64	42	0.84	57	1.35
13	1.16	28	1.68	43	1.57	58	1.05
14	0.35	29	1.46	44	0.84	59	1.10
15	1.31	30	1.41	45	2.54	60	0.90

Table 4 shows that two (2)items, that is, Items 2 and 7 are within the range of 0.00 to 0.34 indicating very low discriminating power, four (4)items, that is, Items 3, 4, 14 and 23 are within the range of 0.35 to 0.64, indicating low discriminating power, eighteen (18) items, that is, Items 6, 9, 11, 13, 15, 16, 20, 21, 25, 26, 37, 40, 41, 42, 44, 58, 59 and 60 are within the range of 0.65 to 1.34, indicating moderate discriminating power, nineteen (19) items, that is, Items 1, 5, 8, 10, 12, 19, 22, 28, 29, 30, 31, 35, 38, 43, 48, 51, 53, 56 and 57 are within the range of 1.35 to 1.69, indicating high discriminating power and seventeen (17) items, that is, Items 17, 18, 24, 27, 32, 33, 34. 36, 39, 45, 46, 47, 49, 50, 52, 54 and 55 are above 1.70 indicating very high discriminating power.

Summary of Findings

1. The 2020 NECO Mathematics multiple-choice test items are multidimensional and locally independent.
2. The 4-parameter logistics IRT model was the best fit for the 2020 NECO Mathematics multiple-choice test items.
3. The difficulty indices of twenty eight (28) items were within the range value of -3 to 0 on the logit scale indicating moderately easy items, while those of twenty nine (29) items were within the range value of 0 to +3 on the logit scale indicating moderately difficult items, and three (3) items were below -3, indicating very easy items under the IRT framework.
4. The discriminating indices of two (2) items were within the range of 0.00 to 0.34 indicating very low discrimination, while twenty-two (22) items within the

range of 0.35 to 1.34 indicated moderate discrimination and thirty-six (36) items with values above 1.35 indicated high discrimination.

Discussion of Findings

Based on Research Question 1, it was found that the 2020 NECO mathematics multiple-choice test items were multidimensional. In other words, different aspects of mathematical abilities were measured by the instrument. The finding agreed with the study of Anyawale, Isaac-Oloniyo and Abayomi (2020) [3] on assessment of dimensionality of Osun State Unified Mathematics Achievement Test items. Findings from the work of Oguoma, Metibemu and Okoye (2016) on dimensionality assumption test on 2014 Mathematics achievement items of West African Senior Secondary Certificate Examination (WASSCE) also indicated that the test items of WASSCE mathematics were inherently multidimensional in nature. Furthermore, Okwilagwe and Ogunrinde (2017) [15] also found that fifty (50) items of 2013 WASSCE and sixty (60) items of National Examinations Council (NECO) Geography violated assumption of unidimensionality and that there were more than one dimension that accounted for the variation observed in examinees to the geography test items.

In terms of local independence, it was found that the assumption was not violated using Yen Q3 statistic. The finding also agreed with Ubi (2006) [20] study that assessed 800 scripts of candidates who wrote Mathematics in the University Matriculation Examination in Cross River State, Nigeria and discovered that a good number of items in the examination were locally independent.

The analysis of model fit was addressed by subjecting the responses to full information item factor analysis, and comparing information obtained, using Akaike Information Criterion (AIC), and Likelihood Ratio Test (LogLik), It was found out that the four parameter logistic model (4PLM) had the smallest information criteria in terms of Akaike information criteria and -2Loglik, and hence, the best model fit for the data. The decision was based on the view of Thorpe and Favia (2012) [19] that the model with the smallest -2loglik value is the best fit.

The Research Question 3 estimated the item difficulty parameters of the 2020 NECO mathematics multiple-choice questions using the item response theory. In IRT, the theoretical range of values of the item difficulty parameters lie between -∞ and +∞, but the typical values range from -3 to +3. Values outside this usual typical range are rare to come by (Baker 2001) [5]. Positive estimates of item difficulty progressively imply difficult items and negative estimates retrogressively imply easy items. The result showed that twenty-eight (28) items, representing 46.7% of the items, fell between-3 to 0 on the logit scale, indicating moderately easy items, while twenty-nine (29) items, representing 48.3% of the items, fell between 0 to +3 on the logit scale, indicating moderately difficult items. Three (3) items, representing 5% of the items, however, fell below -3 on the logit scale indicating, very easy items. Therefore, under the IRT framework, three (3) items were found to be too easy.

The Research Question 4 estimated the item discrimination parameters of the 2020 NECO mathematics multiple-choice questions using the item response theory. Baker (2001) [5] provided the following bases for interpretation of item discrimination parameters:

0.01 - 0.34: Very low
0.35 - 0.64: Low
0.65 - 1.34: Moderate
1.35 - 1.69: High
1.70 and above: Very high

The result showed that two (2) items, representing 3.3% of the items, were within the range of 0.01 to 0.34, indicating very low discriminating power, four (4) items, representing 6.7% of the items, were within the range of 0.35 to 0.64, indicating low discriminating power, eighteen (18) items, representing 30% of the items, were within the range of 0.65 to 1.34, indicating moderate discriminating power, nineteen (19) items, representing 31.7% of the items, were within the range of 1.35 to 1.69, indicating high discriminating power, and seventeen (17) items, representing 28.3% of the items, were above 1.69, indicating very high discriminating power. This interpretation of the item discrimination parameters was based on the categorization of Baker (2001) [5].

Conclusion

The findings from this study indicated that only three items were too easy and six items had low to very low discriminating powers. This shows that the examination body produced quality test items, and it is highly commendable because as emphasized by Nenty (2004), one of the principal tasks in educational practice is the development of tests that measure the facets of learning with the greatest precision and accuracy, and this is associated with the quality of test items. Additionally, the two item statistics (item difficulty and item discrimination) produced almost the same item characteristics, and either could be used to judge the quality of the items.

Recommendations

The following recommendations are made in the light of the above findings:

1. Item analysis should be maintained in the test development process because of its importance in determining the overall effectiveness of the test.
2. Workshops should be organized to educate teachers on the implications of quality test items and its effect on student's performances.
3. IRT and its related applications should be popularized in Nigeria by creating awareness, sponsoring and procuring the various IRT analytical software.

References

1. Adegoke BA. Comparison of item statistics of physics achievement test using classical test theory and item response theory frameworks. Journal of Education and Practice,2013:4(22):87-90.
2. Anigbo LC. Teachers' handbook on measurement and evaluation for effective teaching and learning. Enugu: Executive Press Limited, 2014.
3. Anyanwale MA, Isaac-Oloniyo, Abayomi FR. Dimensionality assessment of binary response test items. A non-parametric approach of Bayesian item response theory measurement. International Journal of Evaluation and Research in Education,2020:9(2):385-393.

4. Ashikhia DA. Students and teachers' perceptions of the causes of poor Academic performance in Ogun State secondary schools (Nigeria): Implication for counseling for national development, 2010. Retrieved from http://www.eurojournal.com/ejss.
5. Baker F. The basics of item response theory. Eric Clearing House on Assessments and Evaluation. University of Maryland College. Park M.D, 2001.
6. Federal Republic of Nigeria (FRN). National policy on education (6th ed.). Lagos: NERDC press, 2014.
7. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. San Diego, CA: Academic Press, 2000.
8. Hambleton RK, Robin F, Xing D. Item response models for the analysis of educational and psychological test data. In H. Tinsley & S. Brown (Eds.) Handbook of applied multivariate statistics and modelling. San Diego, CA: Academic Press, 2000.
9. Marais I. Local dependence. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), Rasch models in health London, England: Wiley, 2013, 111-130.
10. Mbugua ZK, Kibet K, Muthaa G, Nkonke GR. Factors contributing to students' poor performance in mathematics at Kenya Certificate of Secondary Education in Kenya: A case of Baringo County, Kenya. Retrieved, 2012-2014. from http://www.aijcrnet.com/journals/Vol_2_No_6_June_2012/11.pdf
11. Nenty HJ. Designing Measurement Instruments for Assessment and Research in Education. A Paper for Publication in a Book Series by Akwa Ibom State College of Education Afaha NSIT AKS Nigeria, 2004.
12. Oguoma CC, Metibemu MA, Okoye RO. An assessment of the dimensionality of 2014 West African secondary school examination mathematics objective test scores in Imo State, Nigeria, African J. Theory Pract. Educ. Assessment,2016:4:18-33.
13. Ojerinde DP. Classical test theory (CTT) vs item response theory (IRT): An evaluation of the comparability of item analysis results. A guest lecture presented at the Institute of Education, University of Ibadan, on 23rd May, 2013.
14. Okoye RO. Educational and psychological measurement and evaluation. Awka: Erudition Publishers, 2015.
15. Okwilagwe MA, Ogunrinde EA. Assessment of unidimensionality and local independence of WAEC and NECO 2013 Geography Achievement Tests. African J. Theory Pract. Educ. Assess,2017:5:31-44.
16. Olabode JO, Adeleke JO. Comparative analysis of item local independence of WAEC and NECO 2012 mathematics objectives test items. Journal of Educational Researchers and Evaluators of Nigeria,2015:1(2):182-190.
17. Omorogiuwa OK. An introduction to educational measurement and evaluation. Benin: Perfect Touch Prints, 2010.
18. Onah EN. Effect of multimedia projection on senior secondary students' achievement and interest in sets in Enugu State, Nigeria. Nsukka: Department of Science Education, University of Nigeria, Nsukka, 2015.
19. Thorpe GL, Favia A. Data analysis using item response theory methodology: An introduction to selected programs and applications, 2012. Retrieved from http://digitalcommons.lidrary.umaine.edu/psy_facpub/20
20. Ubi IO. Item local independence, dimensionality and trend candidates' mathematics performance on University Matriculation Examination in Nigeria. Unpublished Ph.D Dissertation, University of Calabar, Nigeria, 2006.
21. Wang T, Hanson A. Development of an item response model that incorporates response time. A paper presented to the annual meeting of the American Education Research Association in Settle, April, 2001.
22. Wiberg M. Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test. Umea: Kluwer Academic Publications, 2004.